Learning to Fuse Disparate Sentences

Micha Elsner and Deepak Santhanam

School of Informatics
University of Edinburgh

Department of Computer Science
Brown University

June 24, 2011
The problem: text-to-text generation

**Input**

The bodies showed signs of torture.

They were left on the side of a highway in Chilpancingo, in the southern state of Guerrero, state police said.

**Output**

The bodies of the men, which showed signs of torture, were left on the side of a highway in Chilpancingo, state police told Reuters.
Humans fuse sentences:

- Multidocument summaries (Banko+Vanderwende ‘04)
- Single document summaries (Jing+McKeown ‘99)
- Editing (this study)
Motivation

Humans fuse sentences:
- Multidocument summaries (Banko+Vanderwende ‘04)
- Single document summaries (Jing+McKeown ‘99)
- Editing (this study)

Previous work: multidocument case:
- Similar sentences (*themes*)
- Goal: summarize common information
  (Barzilay+McKeown ‘05), (Krahmer+Marsi ‘05), (Filippova+Strube ‘08)…
Our task setting

Sentences fused by professional editors—Related by discourse, but...

Content is not usually similar!
Our task setting

Sentences fused by professional editors—Related by discourse, but...

Content is not usually similar!

Editing data:
- Naturally occurring dataset
- Probably more similar to single-document summary
- Poses problems for standard approaches
Generic framework for sentence fusion

NLP is often fun
NLP is useful

parsing

root NLP is often fun
root NLP is useful

merging

root NLP is often fun
useful

selection

root NLP is useful

fun

linearization/read-out

NLP is fun and useful
Issues with the generic framework

Selection
What content do we keep?
  ▶ Convey the editor’s desired information
  ▶ Remain grammatical

Merging
Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!

Learning
Can we learn to imitate human performance?
Issues with the generic framework

**Selection**

What content do we keep?
- Convey the editor’s desired information
  - Requires discourse; not going to address
  - Remain grammatical

**Merging**

Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!

**Learning**

Can we learn to imitate human performance?
# Issues with the generic framework

## Selection

What content do we keep?
- Convey the editor’s desired information
  - Requires discourse; not going to address
- Remain grammatical
  - **Constraint satisfaction** (Filippova+Strube ‘08)

## Merging

Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!

## Learning

Can we learn to imitate human performance?
Issues with the generic framework

**Selection**
What content do we keep?
- Convey the editor’s desired information
  - Requires discourse; not going to address
  - Remain grammatical
    - **Constraint satisfaction** (Filippova+Strube ‘08)

**Merging**
Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!
**Contribution:** Solve jointly with selection

**Learning**
Can we learn to imitate human performance?
Issues with the generic framework

Selection
What content do we keep?
- Convey the editor’s desired information
  - Requires discourse; not going to address
  - Remain grammatical
    - Constraint satisfaction (Filippova+Strube ‘08)

Merging
Which nodes in the graph match?
Dissimilar sentences: correspondences are noisy!
Contribution: Solve jointly with selection

Learning
Can we learn to imitate human performance?
Contribution: Use structured learning
Overview

Motivation

Setting up the problem

Fusion as optimization
  Jointly finding correspondences
  Staying grammatical

Learning to fuse
  Defining an objective
  Structured learning

Evaluation
Overview

Motivation

Setting up the problem

Fusion as optimization
  Jointly finding correspondences
  Staying grammatical

Learning to fuse
  Defining an objective
  Structured learning

Evaluation
The data

500 article pairs processed by professional editors:

- Novel dataset courtesy of Thomson Reuters

Each article in two versions: *original* and *edited*

We align originals with edited versions to find:
- 175 split sentences
- 132 merged sentences
- ... treat both as fusion examples
The content selection problem

Which content to select?
(Daume+Marcu ‘04), (Krahmer+al ‘08)

Fake content selection
Use simple dynamic programming to align input with truth...
Provide true alignments to both system and human judges.

Input
Uribe appeared unstoppable after the rescue of Betancourt.
His popularity shot to over 90 percent, but since then news has been bad.

True output
Uribe appeared unstoppable and his popularity shot to over 90 percent.
Overview

Motivation

Setting up the problem

**Fusion as optimization**
- Jointly finding correspondences
- Staying grammatical

**Learning to fuse**
- Defining an objective
- Structured learning

Evaluation
Merging dependency graphs

Previous:

Merge nodes deterministically:

- Lexical similarity
- Local syntax tree similarity

For disparate sentences, these features are noisy!
Merging dependency graphs

Previous:

Merge nodes deterministically:
- Lexical similarity
- Local syntax tree similarity
For disparate sentences, these features are noisy!

Our work:

Soft merging: add **merge arcs** to graph
System decides whether to use or not!
Add relative clause arcs between subjects and verbs
(Alternates “police said” / “police, who said”)

- bodies
- showed
- root
- signs
- torture
- side
- highway
- chilpancingo
- north
- hour
- resort
- acapulco
- state
- police
- said
- root
- they
- left
- were
Merging/selection

A fused tree: a set of arcs to keep/exclude

“State police said the bodies were left by the side of a highway”
Not every set of selected arcs is valid...

- **Unconnected fragment**
  - showed
  - signs → torture

- **Cycle**
  - bodies ← showed
  - merged?
  - they ← left

- **Merged node with two heads**
  - bodies ← showed

- **Missing argument (subject)**
  - state → police
  - left
  - said
  - obj
  - sbj
  - rel
Solving with ILP

**Integer Linear Programming (ILP)**

Generic method for hard combinatorial problems
Well-studied practical solutions *(Ilog CPLEX)*

Our ILP based on *(Filippova+Strube ‘08)*, generalized for soft merging...
Similar setup for sentence compression *(Clarke+Lapata ‘08)*

Very efficient for this size problem
Overview

Motivation

Setting up the problem

Fusion as optimization
  Jointly finding correspondences
  Staying grammatical

Learning to fuse
  Defining an objective
  Structured learning

Evaluation
How to fuse?

ILP tells us what fusions are allowed...
Weights in objective tell us which ones are good.

Recipe for structured learning, (Collins ‘02), others:
Measure similarity to the editor’s sentence...

- Not just lexically (the editor can paraphrase, we can’t!)

Look at **connections between** the retained content

**bodies of the men, which** showed signs of torture

were left on the side of a highway...

state police told Reuters **root**
Finding the oracle

Match this structure:

On this graph:

our graph

editor's sentence
Optimizing

- **Loss function/oracle:**
  - Just shown; connections between regions
- **Features:**
  - POS tags, dependency labels, lexical resources
- **Update rule:**
  - Averaged perceptron
Overview

Motivation

Setting up the problem

Fusion as optimization
  Jointly finding correspondences
  Staying grammatical

Learning to fuse
  Defining an objective
  Structured learning

Evaluation
Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations
Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

**Human**

The editor’s fused sentence
Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

<table>
<thead>
<tr>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>The editor’s fused sentence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Readability upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our parsing and linearization on the editor’s sentence</td>
</tr>
</tbody>
</table>
### Human evaluation

Evaluated for **readability** and **content** by human judges:

- 92 test sentences; 12 judges, 1062 observations

<table>
<thead>
<tr>
<th>Human</th>
<th>The editor’s fused sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readability upper bound</td>
<td>Our parsing and linearization on the editor’s sentence</td>
</tr>
<tr>
<td>“and”-splice</td>
<td>All input sentences, spliced with the word “and”</td>
</tr>
</tbody>
</table>
Human evaluation

Evaluated for **readability** and **content** by human judges:

92 test sentences; 12 judges, 1062 observations

<table>
<thead>
<tr>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>The editor’s fused sentence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Readability upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our parsing and linearization on the editor’s sentence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>“and”-splice</th>
</tr>
</thead>
<tbody>
<tr>
<td>All input sentences, spliced with the word “and”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our system output</td>
</tr>
<tr>
<td>Only abstractive system we tested</td>
</tr>
</tbody>
</table>
## Readability

<table>
<thead>
<tr>
<th>System</th>
<th>Avg</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Editor</td>
<td>4.6</td>
<td></td>
</tr>
<tr>
<td>Readability UB</td>
<td>4.0</td>
<td>▶ Poor linearization: gap of .6</td>
</tr>
<tr>
<td>“And”-splice</td>
<td>3.7</td>
<td>▶ System: additional loss of .9</td>
</tr>
<tr>
<td>System</td>
<td>3.1</td>
<td>▶ Average system score still 3, “fair”</td>
</tr>
</tbody>
</table>

- Avg: Average score
- UB: Upper Bound
<table>
<thead>
<tr>
<th>System</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Editor</td>
<td>4.6</td>
</tr>
<tr>
<td>Readability UB</td>
<td>4.3</td>
</tr>
<tr>
<td>“And”-splice</td>
<td>3.8</td>
</tr>
<tr>
<td>System</td>
<td>3.8</td>
</tr>
</tbody>
</table>

- Score close to 4, “good”
Comparison with “and”-splice

“and”-splice content scores comparable to ours, but...

- Spliced sentences too long
  - 49 words vs human 34, system 33
- Our system has more extreme scores

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>“And”-splice</td>
<td>3</td>
<td>43</td>
<td>60</td>
<td>57</td>
<td>103</td>
<td>266</td>
</tr>
<tr>
<td>System</td>
<td>24</td>
<td>24</td>
<td>39</td>
<td>58</td>
<td>115</td>
<td>260</td>
</tr>
</tbody>
</table>
The bodies showed signs of torture.

They were left on the side of a highway in Chilpancingo, in the southern state of Guerrero, state police said.
The suit claims the company helped fly terrorism suspects abroad to secret prisons.

Holder’s review was disclosed the same day as Justice Department lawyers repeated a Bush administration state-secret claim in a lawsuit against a Boeing Co unit.
Our system

Biden a veteran Democratic senator from Delaware that Vice president-elect and Joe had contacted to lobby was quoted by the Huffington Post as saying Obama had made a mistake by not consulting Feinstein on the Panetta choice.

Better parsing/linearization

Vice President-elect Joe Biden, a veteran Democratic senator from Delaware who had contacted...
The White House that took when Israel invaded Lebanon in 2006 showed no signs of preparing to call for restraint by Israel and the stance echoed of the position.
Conclusion

- Naturally occurring data
- Find correspondences jointly with selection
- Supervised structured learning
Future work

New data:
- Classic similar-sentence fusion (McKeown ‘10 corpus)
- Single-document summary

Better techniques:
- Paraphrasing rules

Code available
bitbucket.org/melsner/sentencefusion
Acknowledgements

Thompson-Reuters: Alan Elsner, Howard Goller, Thomas Kim
BLLIP labmates: Eugene Charniak, Stu Black, Rebecca
Mason, Ben Swanson
Funds: Google Fellowship for NLP
All of you!